Accounting for Errors in Data Improves Divergence Time Estimates in Single-cell Cancer Evolution

Kylie Chen,*[1] Jiří C. Moravec,[2,3] Alex Gavryushkin,[3] David Welch,[1] and Alexei J. Drummond[1,4]

[1]School of Computer Science, University of Auckland, Auckland, New Zealand
[2]Department of Computer Science, University of Otago, Dunedin, New Zealand
[3]School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
[4]School of Biological Sciences, University of Auckland, Auckland, New Zealand

***Corresponding author:** E-mail: kche309@aucklanduni.ac.nz.

Associate editor: Rasmus Nielsen

Abstract

Single-cell sequencing provides a new way to explore the evolutionary history of cells. Compared to traditional bulk sequencing, where a population of heterogeneous cells is pooled to form a single observation, single-cell sequencing isolates and amplifies genetic material from individual cells, thereby preserving the information about the origin of the sequences. However, single-cell data are more error-prone than bulk sequencing data due to the limited genomic material available per cell. Here, we present error and mutation models for evolutionary inference of single-cell data within a mature and extensible Bayesian framework, BEAST2. Our framework enables integration with biologically informative models such as relaxed molecular clocks and population dynamic models. Our simulations show that modeling errors increase the accuracy of relative divergence times and substitution parameters. We reconstruct the phylogenetic history of a colorectal cancer patient and a healthy patient from single-cell DNA sequencing data. We find that the estimated times of terminal splitting events are shifted forward in time compared to models which ignore errors. We observed that not accounting for errors can overestimate the phylogenetic diversity in single-cell DNA sequencing data. We estimate that 30–50% of the apparent diversity can be attributed to error. Our work enables a full Bayesian approach capable of accounting for errors in the data within the integrative Bayesian software framework BEAST2.

Key words: **single-cell sequencing, Bayesian methods, phylogenetics.**

Introduction

The growth of cancer cells can be viewed as an evolutionary process where mutations accumulate along cell lineages over time. Within each cell, single-nucleotide variants (SNVs) act as markers for the evolutionary process. By sampling and sequencing cells, we can reconstruct the possible evolutionary histories of these cell lineages. This can provide insight into the timing of events and modes of evolution.

Currently, there are two main methods for obtaining genomic sequences, bulk sequencing, and single-cell sequencing. Bulk sequencing data are traditionally used in genomic studies. By pooling the genetic material from many cells to form a single observation, greater coverage and thus genetic signal is retained. However, in the context of cancer phylogenetics, the analysis of bulk data poses challenges. Firstly, the intermixing of tumor and normal cells affects the genomic signal. Secondly, the pooled sample may be heterogeneous and thus contain a mixture of different genomic variants (de Bruin et al. 2014; Dagogo-Jack and Shaw 2018; Liu et al. 2018).

In contrast, single-cell sequencing isolates and amplifies the genetic material within a single cell (Kuipers et al. 2017a). The isolation step alleviates the mixture problem. However, errors are more problematic for single-cell sequencing due to insufficient coverage caused by the limited amount of genetic material. The main sources of errors in single-cell sequencing include: cell doublets, where two cells are sequenced as one by mistake; allelic dropout (ADO), where one of the alleles fails to be amplified; and sequencing error, where a base is erroneously read as a different base by the sequencing machine (Kuipers et al. 2017a; Woodworth et al. 2017; Lähnemann et al. 2020). Error models proposed to address these issues include models based on false positives and false negatives (Jahn et al. 2016; Ross and Markowetz 2016; Zafar et al. 2017, 2019), models of allelic dropout and sequencing errors (Kozlov et al. 2022), and models of read count errors (Satas et al. 2020).

To enable easy integration with molecular clock and phylogeography models that are commonly used in other areas of phylogenetics (Meijer et al. 2012; Malmstrøm et al. 2016; Kearns et al. 2018), we implemented two error models within a mature Bayesian evolutionary framework, BEAST2 (Bouckaert et al. 2019). Our motivation is to enable inference and quantify the uncertainty of both the evolutionary history and model parameters for single-cell phylogenetics. Our paper implements: (i) a model for false positive and false negative errors (Jahn et al. 2016; Ross and Markowetz 2016; Zafar et al. 2017, 2019) and (ii) a model for ADO and sequencing errors (Kozlov et al. 2022).

We show that our implementation is well calibrated (Dawid 1982) and demonstrate these models on real and simulated single-cell DNA data. Our simulation studies show that not accounting for errors leads to inaccurate estimation of timing and substitution parameters when data is error-prone. Our results suggest that using a model that is not error-aware can significantly overestimate the number of substitutions and hence the evolutionary time scale. Analysis of empirical single-cell datasets suggests 30–50% of the phylogenetic diversity can be attributed to errors. Moreover, we show error models are feasible on real datasets, with no additional runtime costs compared to the equivalent non-error version of these models. Finally, we should note that these methods, while developed with cancer analysis in mind, are also applicable to non-cancer single-cell phylogenetics, such as somatic cell evolution.

Related Work

For bulk sequencing, there are many tools that estimate the clonal compositions in each bulk sample (Popic et al. 2015; Jiang et al. 2016; Miura et al. 2018) and infer their clonal history (Cooper et al. 2015; Alves et al. 2019; Heide et al. 2019). These clone inference tools are most applicable to bulk sequencing samples that contain a mixture of clones, and phylogenetic reconstruction is performed on the identified clones. However, with the recent availability of single-cell technology, variations between cells can be studied more directly (Schwartz and Schäffer 2017). This has led to the development of tools for single-cell phylogenetics.

As errors present a key challenge to the analysis of single-cell data, there is a need for models that account for errors introduced during the sequencing process, missing data, coverage discrepancies (Lee et al. 2020), and the ability to quantify uncertainty (Lähnemann et al. 2020).

Early models are based on false positive and false negative errors where the input is a mutation matrix in binary format as in OncoNEM (Ross and Markowetz 2016), SCITE (Jahn et al. 2016), SiFit (Zafar et al. 2017), or ternary format as in SiFit (Zafar et al. 2017). OncoNEM (Ross and Markowetz 2016) is a maximum-likelihood (ML) method and uses a heuristic search to optimize the likelihood. SCITE (Jahn et al. 2016) is a Bayesian method that uses Markov chain Monte Carlo (MCMC) to sample the posterior but can also be operated in ML mode. Both OncoNEM and SCITE make the infinite sites assumption where a mutation can occur only once at a site. This assumption may be violated on real data, such as by parallel driver mutations (Tarabichi et al. 2021). Besides this, OncoNEM has been shown to be computationally slow, with low phylogenetic accuracy in the presence of ADO (Kozlov et al. 2022). The infSCITE model (Kuipers et al. 2017b) extends SCITE to account for cell doublet errors and test for the infinite sites assumption. SiFit relaxes the infinite sites assumption and additionally accounts for loss of heterozygosity where a single allele is deleted. As deletion events commonly occur across a large region of the chromosome, this could violate the site independence assumption made by the SiFit model.

SCARLET (Satas et al. 2020) implements a read count model which accounts for false positives and false negatives. This is done by correcting read counts at each site using copy-number variation (CNV) output from another software. Empirical studies have shown ADO is the most significant contributor of errors in single-cell DNA sequencing (Wang and Navin 2015). CellPhy (Kozlov et al. 2022) explicitly models both ADO and sequencing error on diploid genotypes. Unlike models based on false positives and false negatives where different error types are absorbed into the false positive and false negative parameters, CellPhy is a more realistic model of the errors arising from the sequencing process. Furthermore, CellPhy has been shown to produce the most accurate phylogenetic estimates, followed by SiFit and infSCITE on simulated NGS datasets with ADO, amplification, and doublet errors (Kozlov et al. 2022). Both SCARLET and CellPhy make important advances in using data that is closer to the observed sequencing data than previous methods.

Besides CellPhy, which uses the ML phylogenetic framework RAxML (Stamatakis 2014), other methods are only available as standalone implementations. The advantage of our work is that it enables easy integration with a wide range of population and clock models. In doing so, making these models available for single-cell phylogenetics. This includes relaxed clock models (Drummond et al. 2006), population growth models such as Bayesian skyline plots (Drummond et al. 2005), and phylogeography models such as structured coalescent (Vaughan et al. 2014) and isolation migration models (Nielsen and Wakeley 2001).

In this paper, we implement error models for binary SNV data and diploid nucleotide SNV data. The binary model accounts for false positive and false negative errors (Jahn et al. 2016; Ross and Markowetz 2016; Zafar et al. 2017, 2019). The nucleotide model accounts for ADO and sequencing errors (Kozlov et al. 2022). First, we investigate how errors impact the time scale of evolutionary trees inferred from single-cell data. Then, we perform preliminary analyses on real single-cell data to show error models can be used with population growth and molecular clock models. The next section describes the evolutionary models used in this study.

Materials and Methods

We implement two sets of models: (i) the binary model, which handles mutation presence–absence data and (ii) the GT16 model, which handles diploid nucleotide genotypes.

The mutation process is modeled as a substitution process evolving along the branches of a tree τ, with mutation rates defined by the substitution rate matrix Q. Errors are modeled as a noisy process on tip sequences of the tree, where the true genotype is obfuscated according to error probabilities. To perform inference on data, we sample the posterior distribution of trees and the model parameters using Markov chain Monte Carlo (MCMC).

Software and Input format

Our software is available at www.github.com/bioDS/beast-phylonco. It accepts input files in Nexus, FASTA, and VCF format via a conversion script available at www.github.com/bioDS/vcf2fasta.

Binary Substitution Model

The presence or the absence of mutation is represented as a binary state $\Gamma = \{1, 0\}$. The rate matrix Q has a single parameter, λ which is the rate of back-mutation $1 \rightarrow 0$, relative to a mutation rate of 1.

The elements of the rate matrix Q are:

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} -1 & 1 \\ \lambda & -\lambda \end{pmatrix} \end{array}.$$

The equilibrium frequencies are:

$$\pi_0 = \lambda/(\lambda + 1),$$
$$\pi_1 = 1/(\lambda + 1).$$

For data sampled at a single time point, the mutation rate is in units of substitutions per site. Data sampled at multiple time points are required to estimate the mutation rate, which typically has units of substitutions per site per year. Alternatively, if we have prior information on the mutation rate, such as from empirical experiments, we can also fix the model's mutation rate to the empirical value.

Binary Error Model

To account for false positive and false negative errors, we implement the binary error model described in (Jahn et al. 2016; Ross and Markowetz 2016; Zafar et al. 2017). Let α be the false positive probability and β be the false negative probability. $P(x \mid y)$ is the conditional error probability of observing noisy data x, given that the true state is y. For the binary error model, these error probabilities are:

$$\begin{aligned} P(0 \mid 0) &= 1 - \alpha, \\ P(1 \mid 0) &= \alpha, \\ P(0 \mid 1) &= \beta, \\ P(1 \mid 1) &= 1 - \beta. \end{aligned} \tag{1}$$

GT16 Substitution Model

To model diploid nucleotide sequences, we implement the GT16 substitution and error model described in Kozlov et al. (2022). The GT16 substitution model is an extension of the four-state general time-reversible nucleotide GTR model (Tavaré 1986) to diploid genotypes: $\Gamma = \{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT\}$.

Let a, b, c, d be alleles chosen from nucleotides $N = \{A, C, G, T\}$ and r_{ab} be the rate of going from allele a to allele b.

The elements of the rate matrix Q are:

$$\begin{aligned} Q_{aa \rightarrow ab} &= r_{ab} \cdot \pi_{ab}, \\ Q_{aa \rightarrow ba} &= r_{ab} \cdot \pi_{ba}, \\ Q_{ab \rightarrow aa} &= r_{ab} \cdot \pi_{aa}, \\ Q_{ab \rightarrow bb} &= r_{ab} \cdot \pi_{bb}. \end{aligned}$$

Other non-diagonal entries not listed above have a rate of zero. The diagonals are the sum of the in-going rates:

$$Q_{aa \rightarrow aa} = - \sum_{b,c \in N \setminus a} Q_{aa \rightarrow bc},$$
$$Q_{ab \rightarrow ab} = - \sum_{c,d \in N \ c \neq a \ or \ d \neq b} Q_{ab \rightarrow cd}.$$

The relative rates of the Q matrix are:

$$\begin{aligned} r_{AC} &= r_{CA} = \alpha, \\ r_{AG} &= r_{GA} = \beta, \\ r_{AT} &= r_{TA} = \gamma, \\ r_{CG} &= r_{GC} = \kappa, \\ r_{CT} &= r_{TC} = \lambda, \\ r_{GT} &= r_{TG} = \mu. \end{aligned}$$

The equilibrium frequencies are: $\pi = (\pi_{AA}, \pi_{AC}, \pi_{AG}, \pi_{AT}, \pi_{CA}, \pi_{CC}, \pi_{CG}, \pi_{CT}, \pi_{GA}, \pi_{GC}, \pi_{GG}, \pi_{GT}, \pi_{TA}, \pi_{TC}, \pi_{TG}, \pi_{TT})$.

GT16 Error Model

The GT16 error model for diploid nucleotides described in (Kozlov et al. 2022) accounts for amplification errors and biases in single-cell sequencing. This model has two parameters, the combined amplification and sequencing error ϵ, and allelic dropout error δ. The error probabilities $P(x \mid y)$ for genotypes with alleles a, b, c derived in Kozlov et al. (2022) are given below:

$$\begin{aligned} P(aa \mid aa) &= 1 - \epsilon + (1/2) \cdot \delta\epsilon, \\ P(ab \mid aa) &= (1 - \delta)(1/6) \cdot \epsilon, \\ P(bb \mid aa) &= (1/6) \cdot \delta\epsilon, \\ P(aa \mid ab) &= (1/2) \cdot \delta + (1/6) \cdot \epsilon - (1/3) \cdot \delta, \\ P(cc \mid ab) &= (1/6) \cdot \delta\epsilon, \\ P(ac \mid ab) &= (1 - \delta)(1/6) \cdot \epsilon, \\ P(ab \mid ab) &= (1 - \delta)(1 - \epsilon). \end{aligned} \tag{2}$$

Here, we assume that $P(ba\,|\,aa) = P(ab\,|\,aa)$ and $P(cb\,|\,ab) = P(ac\,|\,ab)$. Other combinations not listed above have zero probability. These genotypes can be easily adapted to unphased data by encoding heterozygous states as ambiguities $P(ab^*) = P(ab) + P(ba)$, where ab^* represents ab without phasing information. Our implementation can handle both phased and unphased data.

Likelihood Calculation

Our input data is a SNV matrix D with n sites and m cells. The cell evolutionary tree τ is a rooted binary tree with m cells at the leaves, and branch lengths $t_1, t_2, \ldots t_{2(m-1)}$. These branch lengths are scaled to units of substitutions per site for data sampled at a single time point or years where multiple time points are available. Figure 1 shows an example tree with cells a, b, c, d sampled at different time points.

The likelihood $P(D\,|\,\tau, M, \theta)$ is the conditional probability of observing data D, given a tree τ, a substitution model M with rate matrix Q and model parameters θ. Assuming each site i evolves independently, this likelihood can be written as:

$$P(D\,|\,\tau, M, \theta) = \prod_{i=1}^{n} P(D_i\,|\,\tau, M, \theta).$$

This can be calculated using Felsenstein's peeling algorithm (Felsenstein 1981) by recursively traversing the tree. The likelihood at the root node g, $P(D_i\,|\,\tau, M, \theta)$ is calculated by multiplying the equilibrium frequency π_x of genotype x at site i with its partial likelihood $L_i^g(x)$ summed over all possible genotypes $x \in \Gamma$:

$$P(D_i\,|\,\tau, M, \theta) = \sum_{x\in\Gamma} \pi_x \cdot L_i^g(x).$$

The partial likelihood $L_i^g(x)$ for an internal node g at site i with child nodes e and f and corresponding branch lengths t_e and t_f is:

$$L_i^g(x) = \sum_{y\in\Gamma} P_{xy}(t_e) \cdot L_i^e(y) \cdot \sum_{z\in\Gamma} P_{xz}(t_f) \cdot L_i^f(z),$$

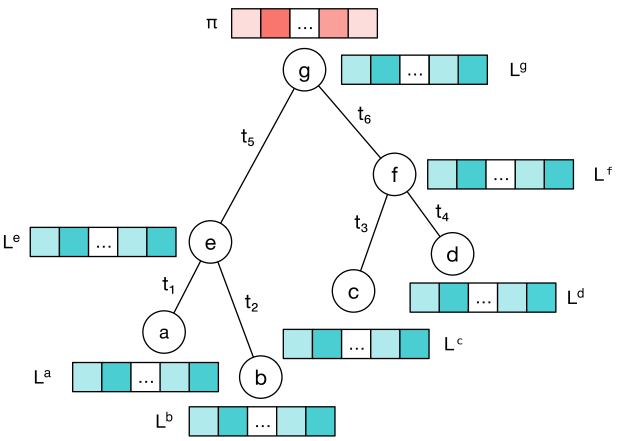

Fig. 1. Example of an evolutionary tree with four cells at the leaves *a*, *b*, *c*, *d*.

where $P_{xy}(t) = [e^{Qt}]_{xy}$ is the probability of going from genotype x to genotype y after branch length t and Q is the rate matrix of the substitution model.

Without an error model, the likelihood vector of a leaf node c with observed genotype x at site i is:

$$L_i^c(x) = \begin{cases} 1 & \text{if } y = x \text{ and } y \in \Gamma, \\ 0 & \text{if } y \neq x \text{ and } y \in \Gamma. \end{cases}$$

To incorporate errors, we replace the leaf likelihood vectors with the conditional error probabilities $P(x\,|\,y)$ in the error model. For example, using an error model the leaf node c with observed genotype x at site i is updated to be:

$$L_i^c(x) = \{P(x\,|\,y), \forall\, y \in \Gamma\}.$$

In the binary error model, the leaf likelihood vector $L_i^c(x)$ for node c is filled using Equation (1) based on its observed genotype x at site i:

$$L_i^c(x) = \begin{cases} (1-\alpha,\ \ \beta) & \text{if } x = 0, \\ (\alpha,\ \ 1-\beta) & \text{if } x = 1. \end{cases}$$

These leaf likelihoods collapse to the non-error version when $\alpha = 0$ and $\beta = 0$. For brevity, we only fully write out the likelihood of the binary model. The likelihood of leaf nodes for the GT16 error model can be derived similarly using equation (2) and is available in our implementation.

Results

Evaluation on Simulated Datasets

First, we evaluated our implementations using a well-calibrated study (Dawid 1982) to test the reliability of the inference when simulating directly from the model. Following the well-calibrated criterion for credible intervals, we expect 95% of the credible interval to cover the true value 95% of the time. Next, we simulated sequences with errors and compared the inference performance with and without modeling the error. Then, we compared the runtime and convergence efficiency of each error model with the baseline non-error substitution model. Lastly, we performed experiments on data simulated with high levels of errors to test the robustness of our methods.

Simulation 1: Binary Data

We performed a well-calibrated study for the binary model using binary sequences with errors. First, we generated trees using a Yule model, then binary sequences were simulated along the branches of the tree, and errors were applied at the tips. Using the sequence data, we jointly estimate the model parameters, tree topology, and branch times using the binary model.

Simulation parameters: We generated 100 trees with 30 leaves from a Yule model, where each tree has a birthrate drawn from Normal($\mu = 7.0$, $\sigma = 1.0$). Sequences of length 400 were simulated using the binary model with rate $\lambda \sim$ Lognormal($\mu = -1$), false positive probability $\alpha \sim$ Beta(1, 50) and false negative probability $\beta \sim$ Beta(1, 50).

Supplementary figure S1, Supplementary Material shows the estimates for the model parameters, tree length, and tree height compared to the true simulated values. The estimated 95% highest posterior density (HPD) intervals are shown as bars, where blue indicates the estimate covers the true value, and red indicates otherwise. Our simulations show that the true value of each parameter falls within the estimated 95% HPD interval 91–99% of the time. Supplementary figure S2, Supplementary Material shows the estimated trees are, on average, 2–5 subtree prune and regraft (SPR) moves away from the true tree.

Simulation 2: Binary Data Error vs. No Error

To compare the effects of inference with and without error modeling, we used the data from Simulation 1, then performed inference with and without the binary error model. We compared the coverage for each parameter with and without an error model, that is, how often the estimated 95% HPD covers the true value.

Supplementary figure S1, Supplementary Material shows the estimated parameters with the binary error model, and supplementary figure S3, Supplementary Material shows the estimated parameters without using an error model. Supplementary table S1, Supplementary Material shows the coverage of each parameter. The coverage of tree length drops from 95% when the error model is used to 39% when no error model is used. Similarly, the coverage of the substitution parameter λ drops from 91% to 53%. Furthermore, the tree length tends to be overestimated when no error model is used. This suggests that both the tree length and substitution parameters are significantly biased when errors present in the data are not modeled. Figure 2 shows a comparison of the estimated tree length and tree height for these two model configurations. Other parameters such as birthrate and tree height are less biased, with a coverage of 92% and 84% respectively when no error model is used.

Simulation 3: Diploid Nucleotide Phased Data

We performed a well-calibrated study for the GT16 model using phased sequence data with errors. First, we simulated trees using a coalescent model. Sequences were simulated down branches of the tree using the GT16 substitution model, and then errors were applied at the tips. Using these sequences as input, we estimated the tree and model parameters using the GT16 model. The priors on the error probabilities were chosen based on experimental studies for allelic dropout (Huang et al. 2015), amplification and sequencing errors (Gawad et al. 2016; Ross et al. 2013).

Simulation parameters: We generated 100 trees with 16 leaves from a coalescent model, where the population size is drawn from $\theta \sim$ LogNormal($\mu = -2.0, \sigma = 1.0$). Sequences of length 200 were simulated using the GT16 model with genotype frequencies $\pi \sim$ Dirichlet(3, 3, ..., 3), and relative rates $r \sim$ Dirichlet(1, 2, 1, 1, 2, 1). Errors were simulated using $\epsilon \sim$ Beta($\alpha = 2, \beta = 18$), and $\delta \sim$ Beta($\alpha = 1.5, \beta = 4.5$).

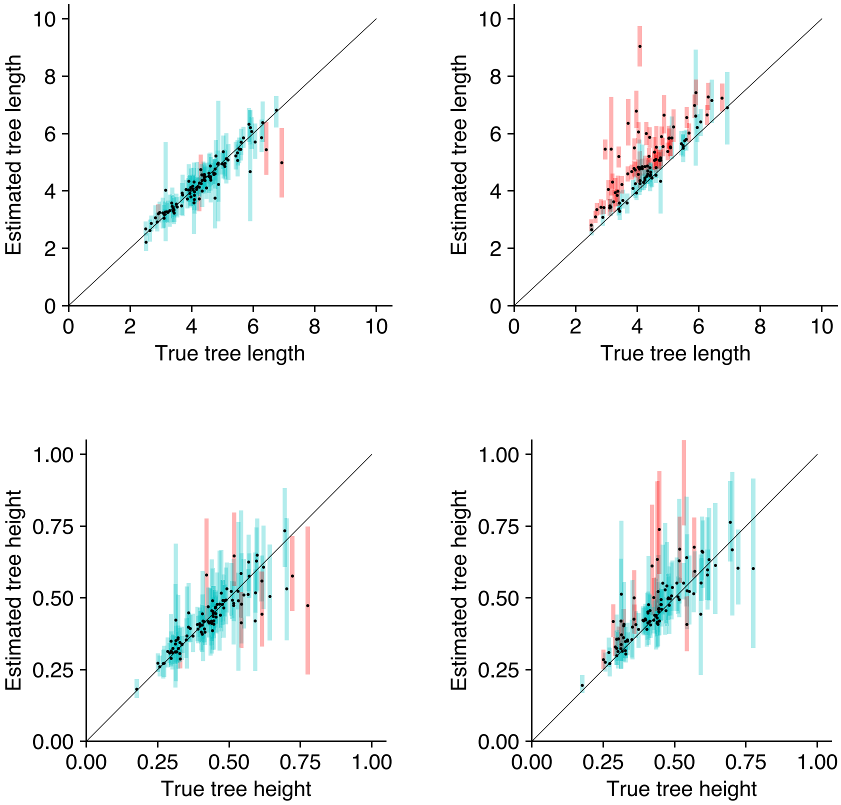

Fig. 2. Comparison of branch lengths with and without the binary error model on simulated binary data. Estimated branch lengths using the binary error model (left) and without using the error model (right). Tree length estimates higher than 10, and tree height estimates higher than 1 are truncated on this plot.

Supplementary figures S4–S9, Supplementary Material show the 95% HPD estimated for each model parameter and tree branch lengths. The true value of each parameter falls within the 95% HPD interval 91–99% of the time. This shows we are able to accurately estimate the substitution, error, population parameters, and branch lengths for phased data. Next, we computed the accuracy of tree topology by comparing the estimated tree with the true tree. Supplementary figure S10, Supplementary Material shows the average distance from the estimated trees to the true tree. On average, estimated trees are 2–6 SPR moves away from the true tree.

Simulation 4: Diploid Nucleotide Unphased Data

We performed a well-calibrated study for the GT16 model using unphased sequencing data. For unphased data, we used the data generated from Simulation 3, with phasing information removed from the sequences. Phasing information was removed by mapping a heterozygous *ab* to both states *ab* and *ba*.

Supplementary figures S11–S15, Supplementary Material show the estimated model parameters compared to the true simulated values. For each parameter, the estimated 95% HPD interval covers the true value 94–99% of the time, which confirms our implementation is well-calibrated. On average, the estimated trees are 2–6 SPR moves away from the true tree, supplementary figure S16, Supplementary Material. We note that the sum of the paired heterozygous frequencies ($\pi_{ab} + \pi_{ba}$) are identifiable, but the individual frequencies (π_{ab}, π_{ba}) are non-identifiable as the data are unphased.

Simulation 5: Diploid Nucleotide Data Error vs. No Error

To compare the effects of inference with and without error modeling for diploid nucleotide data, we used the data from Simulation 3, then performed inference with and without the GT16 error model.

Supplementary table S2, Supplementary Material shows the coverage of each parameter with and without an error model. Supplementary figures S17–S21, Supplementary Material show the estimated model parameters when an error model is not used. We observe a similar trend to Simulation 2, where the tree length and substitution parameters are significantly biased without an error model. Although the tree height estimated without an error model are less biased, the tree lengths are overestimated. These differences in the tree heights and tree lengths are highlighted in figure 3.

Simulation 6: Timing Experiments

We measured the runtime and convergence of the error model compared with the baseline non-error implementation in our framework. Both error models are comparable in computational runtime efficiency with their baseline non-error substitution models. Runtime comparisons are shown in supplementary figures S24 and S25, Supplementary Material. The GT16 model takes approximately an hour to reach convergence on simulated datasets with 20 taxa and 500 sites (convergence is measured as the time till the minimum effective sample size is greater than 200). On a similar-sized dataset, the binary model takes less than five minutes to converge. Timing experiments were done on an Intel Xeon E3-12xx v2 virtual

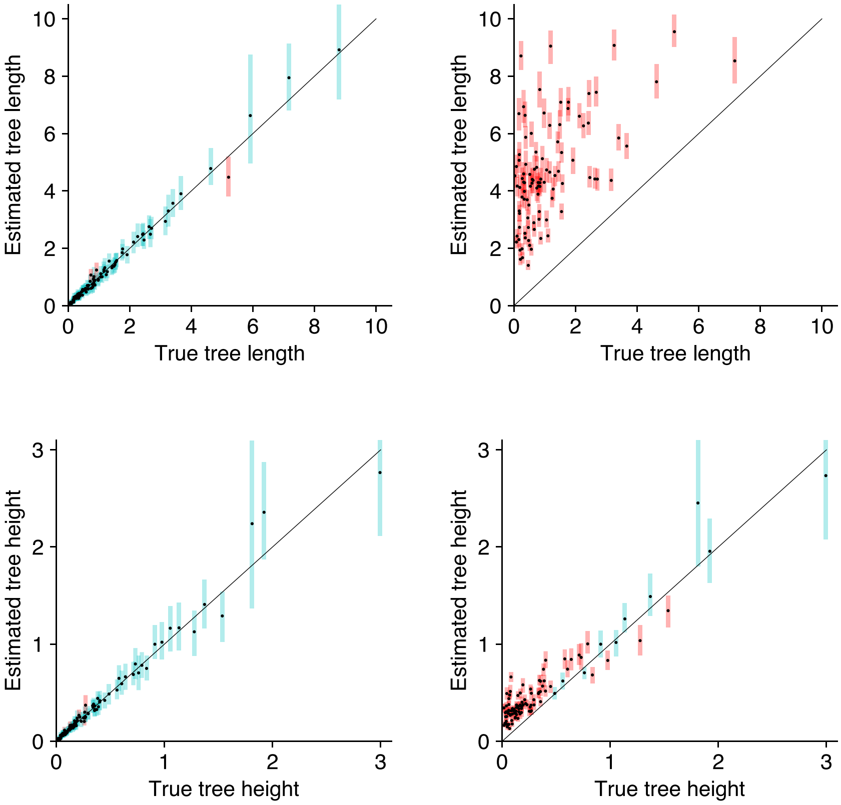

Fig. 3. Comparison of branch lengths with and without the GT16 error model on simulated phased nucleotide data. Estimated branch lengths using the GT16 error model (left) and without using the error model (right). Tree length estimates higher than 10, and tree height estimates higher than 3 are truncated on this plot.

machine with 16 processors at 2.7 MHz and 32 GB RAM hosted by Nectar Research Cloud.

Simulation 7: Performance on Data with High Levels of Error

Lastly, we performed experiments on extended error ranges based on empirical studies (Ross et al. 2013; Huang et al. 2015; Gawad et al. 2016) to test the robustness of our method. We used the same simulation parameters as simulations 1 and 3, but with varying levels of error chosen from an extended range: $\alpha \in [0.001, 0.1]$, $\beta \in [0.1, 0.6]$ for binary data and $\delta \in [0.1, 0.8]$, $\epsilon \in [0.001, 0.1]$ for diploid genotype data. The priors on the error parameters are $\alpha \sim \text{Beta}(1, 20)$ and $\beta \sim \text{Beta}(3, 3)$ for binary data and $\delta \sim \text{Beta}(1.5, 4.5)$ and $\epsilon \sim \text{Beta}(2, 18)$ for diploid nucleotide data. Our results in the supplementary materials, Supplementary Material confirm our methods are robust to high levels of error.

Evaluation on Single-cell Datasets

We analyzed two public datasets from previously published studies; L86, a colorectal cancer dataset (Leung et al. 2017), and E15, a healthy neurons dataset (Evrony et al. 2015). Preprocessed SNVs from CellPhy (Kozlov et al. 2022) were used for both L86 and E15.

Colorectal Cancer Dataset (L86)

L86 contains 86 cells sequenced from a colorectal cancer patient with metastatic spread. The cells were sampled from the primary tumor (colorectal), the secondary metastatic tumor (liver), and matched normal tissue. We used the GT16 model with a relaxed clock to allow for different molecular clock rates in cancer and non-cancer lineages and a coalescent skyline tree prior, which allows changes in population sizes through time.

Model parameters: We used a GT16 substitution model with priors of frequencies $\pi \sim \text{Dirichlet}(3, 3, \ldots, 3)$, relative rates $r \sim \text{Dirichlet}(1, 2, 1, 1, 2, 1)$, and GT16 error model with allelic dropout $\delta \sim \text{Beta}(1.5, 4.5)$ and sequencing error $\epsilon \sim \text{Beta}(2, 18)$. A relaxed clock with a Lognormal prior, and Skyline coalescent tree prior with $\theta_1 \sim \text{Lognormal}(\mu = -2.3, \sigma = 1.8)$. We performed two independent repeats of the MCMC chains.

We found that the tree height is similar for both error and non-error models, but the relative ages of terminal branches are shorter for the error model. Figure 4 shows the tree length, treeness (Lanyon 1988; Phillips and Penny 2003), and gamma statistics (Pybus and Harvey 2000) of the tree distributions under different experimental setups: with and without error modeling, and with and without an outgroup constraint. The trees estimated using an error model are more tree-like than ones estimated without an error model. For the default setup without an outgroup, the 95% HPD estimate for tree length is (4.79, 5.85) with the error model and (7.00, 8.19) without the error model. The error parameters estimates are $\delta \sim (0.62, 0.66)$ and $\epsilon \sim (7 \cdot 10^{-6}, 1 \cdot 10^{-3})$. The error estimates are comparable to the estimates reported by CellPhy (Kozlov et al. 2022) which are $\delta \sim 0.63$ and $\epsilon \sim 0.00$.

Figure 5 summarizes the estimated tree with the error model (top) compared to without an error model (bottom). The tips of the tree are colored by cell type. The trees show most cells group together by their cell type, which suggests there is signal in the data. However, there is some intermixing of metastatic tumor cells inside the primary tumor clade and missorted normal cells as previously identified by Leung et al. (2017) and Kozlov et al. (2022). We also note that the most recent common ancestor (MRCA) of the normal clade is younger than the MRCA of the two tumor clades for both analyses. This is not what we intuitively expected because we believe the normal ancestral cell should be the ancestor of both tumor and normal cells. Although surprising, this observation is in agreement with the trees estimated by ML algorithms in Kozlov et al. (2022). We believe this issue is closely related to the phylogenetic rooting problem for heterogeneous data (Tian and Kubatko 2017). Methods by Tian and Kubatko (2017), Mai et al. (2017) and Drummond et al. (2006) have provided some partial solutions to the rooting problem, but further research efforts are required to better understand the effects on tree topology.

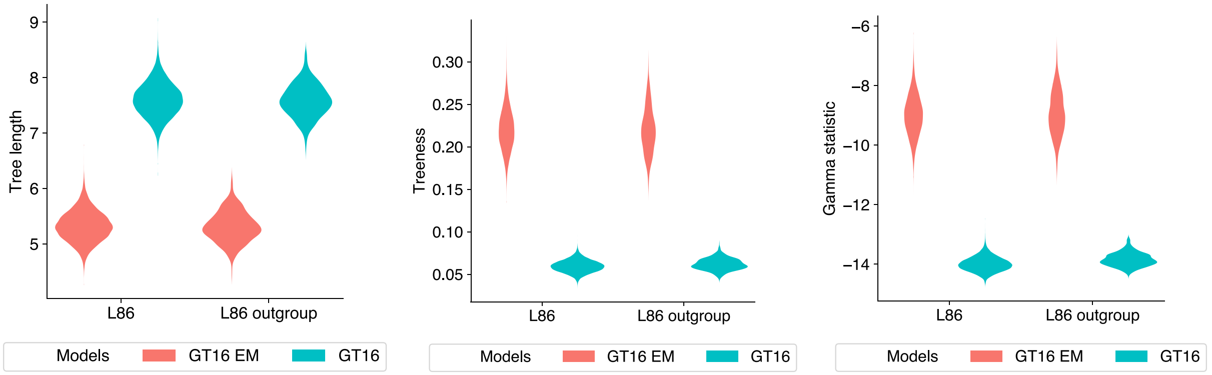

Fig. 4. Tree length, treeness, and gamma statistics of tree distributions estimated from the L86 dataset. The distributions of each metric is colored by the model used: GT16 error model (red) and GT16 model without error (blue). Two pairs of experiments are shown; L86, which has no tree topology constraints, and L86 outgroup, which has the tree topology constrained to normal cells in the outgroup.

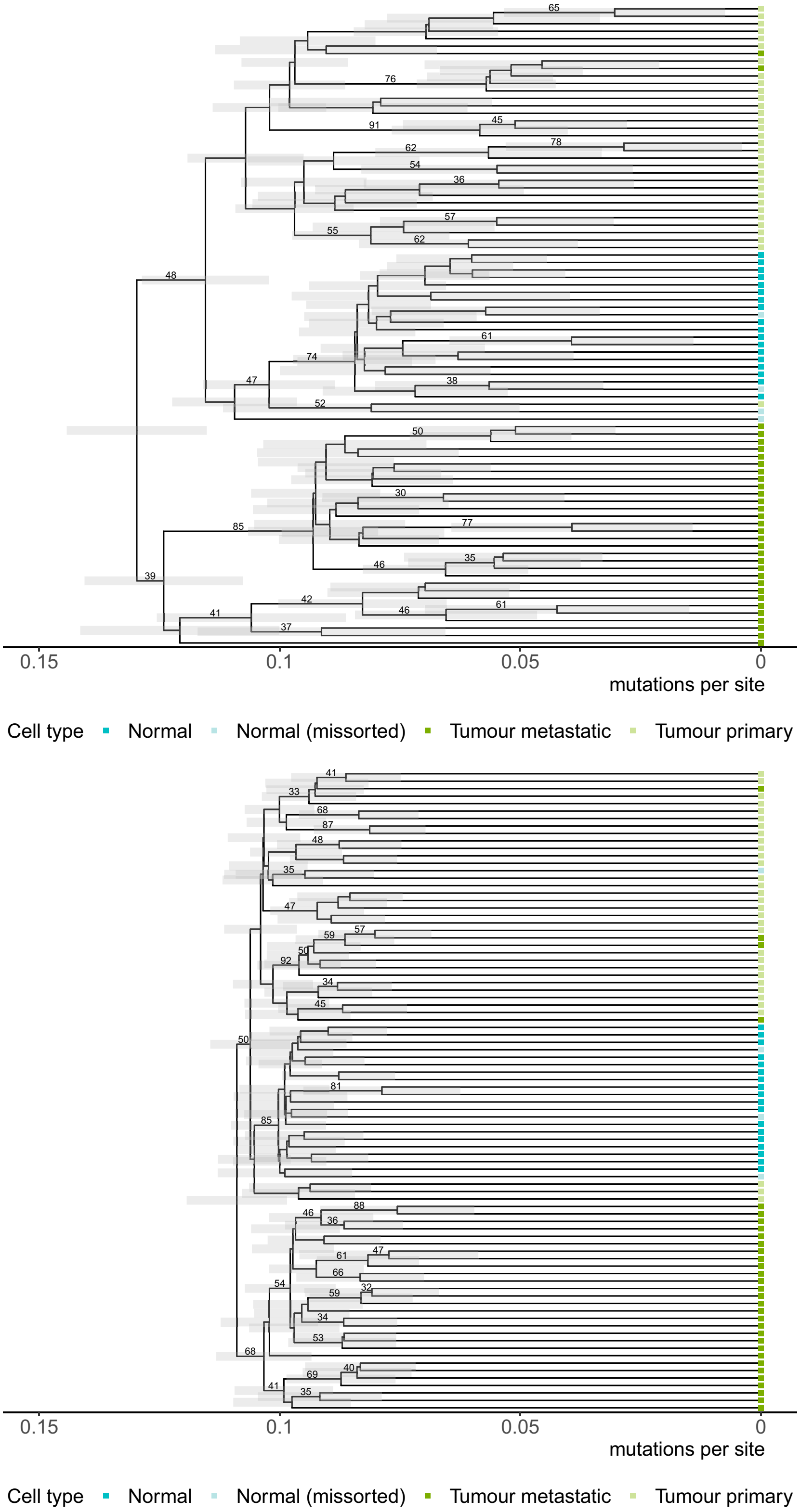

Fig. 5. Maximum clade credibility trees for the L86 dataset (colorectal cancer patient) using the GT16 model with an error model (top) and without an error model (bottom). Cells are colored by their cell types: normal cells (dark blue), normal cells missorted during data collection (light blue), the metastatic tumor from the liver (dark green), and primary tumor from the colon (light green). Clades with greater than 30% are labeled.

We also investigated whether constraining the tree topology to have all tumor cells as the ingroup produces different estimates. We repeat the analyses with an outgroup constraint to the tree topology, setting the normal and missorted samples as the outgroup. Supplementary figure S22, Supplementary Material shows the outgroup constrained summary tree for L86. In this outgroup constrained tree, the age of the MRCA of the normal group is also younger than the MRCA of the primary or metastatic tumors for the error model. The age of the MRCA of the normal group is indistinguishable from the MRCA of the primary or metastatic tumors for the model without error.

Healthy Neurons Dataset (E15)

E15 contains 15 neurons and a blood cell taken from the heart region sequenced from a healthy patient.

Model parameters: We used a GT16 substitution with frequencies prior $\pi \sim$ Dirichlet(3, 3, …, 3) and relative rates $r \sim$ Dirichlet(1, 2, 1, 1, 2, 1), GT16 error model with allelic dropout error $\delta \sim$ Beta(1.5, 4.5), and sequencing error $\epsilon \sim$ Beta(2, 18), with a relaxed clock and Skyline coalescent tree prior with $\theta_1 \sim \text{Lognormal}(\mu = -2.3, \sigma = 1.8)$. We performed two independent repeats of the MCMC chains.

We observed that the trees estimated using the error model are more tree-like than ones estimated without an error model, as shown by the tree metrics in figure 6. The 95% HPD estimates of tree length are (1.37, 7.14) with the error model, and (7.60, 13.41) without the error model. We note the tree height for the error model (0.20, 0.80) is lower than that of the non-error model (0.54, 1.01). The estimated interval for the error parameters are $\delta \sim$ (0.86, 0.92) and $\epsilon \sim$ (0.03, 0.17). To test the sensitivity of the error priors, we reran our experiments with adjusted priors $\epsilon, \delta \sim$ Beta(1, 10) and $\epsilon, \delta \sim$ Beta(1, 20). We found the error estimates were similar regardless of these adjustments on the error parameter priors.

Figure 7 shows a summary of the estimated trees with the GT16 error model (top) and without an error model (bottom). The tips of the tree are colored by cell types. We expect the blood cell to be placed as an outgroup; however, the estimated trees placed the blood cell inside a clade of neuron cells. To investigate if adding an outgroup constraint to the tree topology can help direct the likelihood in the correct direction, we repeated the analyses with the blood cell as the outgroup. The estimated trees are shown in supplementary figure S23, Supplementary Material. Besides the correct placement of the outgroup enforced by the outgroup constraint, we did not observe any substantial topological discrepancies between the outgroup and non-outgroup analyses.

Discussion

We demonstrated that incorporating error parameters can affect the relative ages of single-cell datasets. We showed that models incorporating sequencing error could increase the accuracy of tree branches and model parameters inferred from noisy data. Additionally, we find that using error models is just as fast as the baseline non-error substitution models in our framework. Future work to support multi-threading and add compatibility with the Beagle high-performance library (Ayres et al. 2012) would further increase the computational speed of these models.

From both simulated and real single-cell data, we observed that using an error model tends to shorten the total tree length, as errors explain a portion of the genetic variability within the data. For empirical single-cell data, cells of the same type tend to be placed in the same clade. We believe relaxed clock and local clock models are more suited to heterogeneous data as they allow for changes in mutation rates. The datasets we explored in this paper are sampled at a single time point, so there is no calibration information to allow the mutation rate and time to be disambiguated. Using time sampled data or empirical mutation rate calibrations would improve current analyses and allow node ages to be converted to real time (Drummond et al. 2002, 2003).

Although the effect of filtering strategies in the context of macroevolution shows stringent filtering of sites often leads to worse phylogenetic inference (Tan et al. 2015). The effect of filtering strategies on noisy data such as single-cell phylogenies is yet to be systematically explored. We believe the error parameters in these models can provide increased flexibility, allowing key features of the sequencing and filtering process to be accounted for during evolutionary inference.

Lastly, incorporating cell biology knowledge during method development would improve the biological significance of model assumptions; and improving the interpretability of tree summarization metrics would enable single-cell phylogenies to be examined in more detail.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution online.*

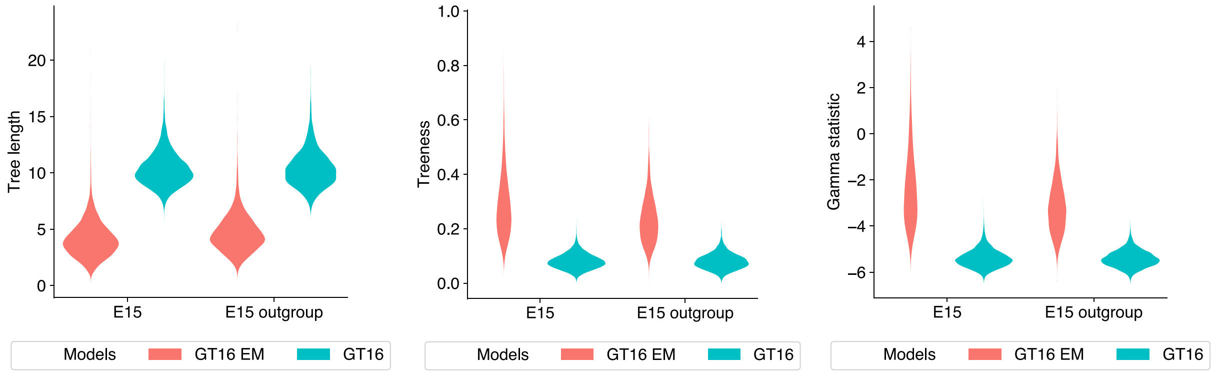

Fig. 6. Tree length, treeness, and gamma statistics of tree distributions estimated from the E15 dataset. The distributions of each metric is colored by the model used: GT16 error model (red) and GT16 model without error (blue). Two pairs of experiments are shown; E15, which has no tree topology constraints, and E15 outgroup, which has the tree topology constrained with the heart cell as the outgroup.

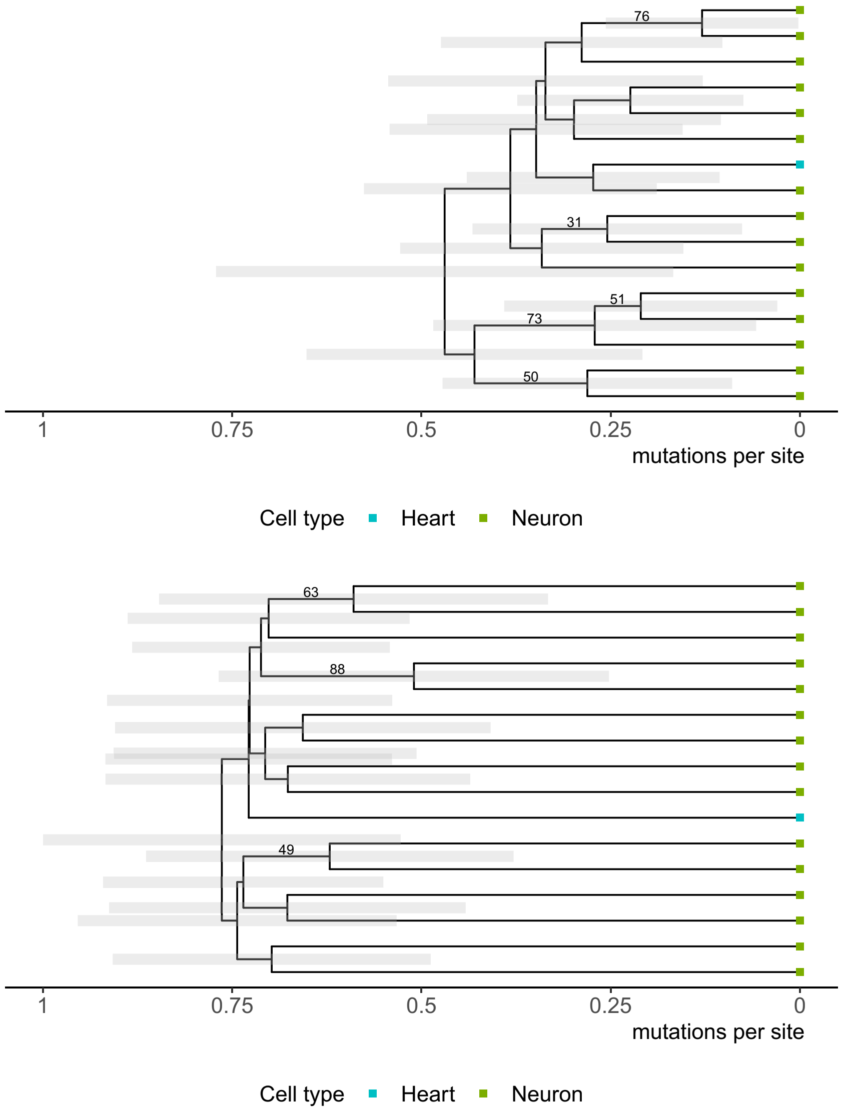

FIG. 7. Maximum clade credibility trees for the E15 dataset (healthy patient) using the GT16 model with an error model (top) and without an error model (bottom). Each cell is colored by its cell type: blood cell (blue) and neuron cells (green). The posterior clade support for clades with greater than 30% support are shown on the branches.

Acknowledgments

We thank Prof. David Posada and Dr. Joao Alves for generously providing us with single-cell datasets and advice. We thank Dr. Remco Bouckaert for helpful discussions and Dr. Walter Xie for implementation support. We acknowledge Nectar Research Cloud and Otago Computing Services for providing the computing resources necessary for this study. We thank the reviewers and editors for their valuable comments which greatly improved the quality of our manuscript.

AJD and AG acknowledge support from a Data Science Programme grant (UOAX1932); AJD, AG and JCM acknowledge support from an Endeavour Smart Ideas grant (U00X1912); AG was supported by a Rutherford Discovery Fellowship (RDF-UOO1702); AJD was supported by a James Cook Research Fellowship from the Royal Society of New Zealand; KC was supported by a University of Auckland Doctoral Scholarship.

Data Availability

Our software, Phylonco v0.0.6 is available at www.github.com/bioDS/beast-phylonco. Analyses, scripts, and data are available at www.github.com/bioDS/beast-phylonco-paper. This paper uses BEAST v2.6.6 (Bouckaert et al. 2019), BeastLabs v1.9.7, LPhy v1.2.0, and LPhyBeast v0.3.0. The following python packages were used: DendroPy (Sukumaran and Holder 2010), lxml (Behnel et al. 2005), matplotlib (Hunter 2007), numpy (Harris et al. 2020), and seaborn (Waskom 2021). The following R packages were used: ggtree (Yu et al. 2017), ggplot2 (Wickham 2016), tracerR (Rambaut et al. 2018), treeSimGM (Hagen and Stadler 2018), treeio (Wang et al. 2020), expm, and ape (Paradis et al. 2019).

References

Alves JM, Prado-Lopez S, Cameselle-Teijeiro JM, Posada D. 2019. Rapid evolution and biogeographic spread in a colorectal cancer. *Nat Commun*. **10**(1):1–7.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, *et al.* 2012. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol*. **61**(1):170–173.

Behnel S, Faassen M, Bicking I. 2005. lxml: Xml and html with python.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, *et al.* 2019. Beast 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLOS Comput Biol*. **15**(4):1–28. doi:10.1371/journal.pcbi.1006650

Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, Kremeyer B, Butler A, Lynch AG, Camacho N, *et al.* 2015. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nat Genet*. **47**(4): 367–372.

Dagogo-Jack I, Shaw AT. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol*. **15**(2):81.

Dawid AP. 1982. The well-calibrated Bayesian. *J Am Stat Assoc*. **77**-(379):605–610.

de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, Jamal-Hanjani M, Shafi S, Murugaesu N, Rowan AJ, *et al.* 2014. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**(6206):251–256.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. **4**(5):e88.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**(3):1307–1320. doi:10.1093/genetics/161.3.1307

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. 2003. Measurably evolving populations. *Trends Ecol Evol*. **18**(9): 481–488. doi:10.1016/S0169-5347(03)00216-7

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. **22**(5):1185–1192.

Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, Yang L, Haseley P, Lehmann HS, Park PJ, *et al.* 2015. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**(1): 49–59.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. **17**(6):368–376.

Gawad C, Koh W, Quake SR. 2016. Single-cell genome sequencing: current state of the science. *Nat Rev Genet*. **17**(3):175.

Hagen O, Stadler T. 2018. Treesimgm: simulating phylogenetic trees under general Bellman–Harris models with lineage-specific shifts of speciation and extinction in R. *Methods Ecol Evol*. **9**(3):754–760.

Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al.* 2020. Array programming with NumPy. *Nature* **585**(7825):357–362.

Heide T, Maurer A, Eipel M, Knoll K, Geelvink M, Veeck J, Knuechel R, van Essen J, Stoehr R, Hartmann A, *et al.* 2019. Multiregion human bladder cancer sequencing reveals tumour evolution, bladder cancer phenotypes and implications for targeted therapy. *J Pathol*. **248**(2):230–242.

Huang L, Ma F, Chapman A, Lu S, Xie XS. 2015. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu Rev Genomics Human Genet*. **16**:79–102.

Hunter JD. 2007. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. **9**(3):90–95. doi:10.1109/MCSE.2007.55

Jahn K, Kuipers J, Beerenwinkel N. 2016. Tree inference for single-cell data. *Genome Biol*. **17**(1):86.

Jiang Y, Qiu Y, Minn AJ, Zhang NR. 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proc Natl Acad Sci*. **113**(37):E5528–E5537.

Kearns AM, Restani M, Szabo I, Schrøder-Nielsen A, Kim JA, Richardson HM, Marzluff JM, Fleischer RC, Johnsen A, Omland KE. 2018. Genomic evidence of speciation reversal in ravens. *Nat Commun*. **9**(1):1–13.

Kozlov A, Alves JM, Stamatakis A, Posada D. 2022. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. *Genome Biol*. **23**(1):1–30.

Kuipers J, Jahn K, Beerenwinkel N. 2017a. Advances in understanding tumour evolution through single-cell sequencing. *Biochim Biophys Acta (BBA)-Rev Cancer*. **1867**(2):127–138.

Kuipers J, Jahn K, Raphael BJ, Beerenwinkel N. 2017b. Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res*. **27**(11):1885–1894.

Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, Vallejos CA, Campbell KR, Beerenwinkel N, Mahfouz A, *et al.* 2020. Eleven grand challenges in single-cell data science. *Genome Biol*. **21**(1):1–35.

Lanyon SM. 1988. The stochastic mode of molecular evolution: what consequences for systematic investigations? *The Auk*. **105**(3): 565–573.

Lee J, Hyeon DY, Hwang D, 2020. Single-cell multiomics: technologies and data analysis methods. *Exp Mol Med*. **52**(9):1428–1442.

Leung ML, Davis A, Gao R, Casasent A, Wang Y, Sei E, Vilar E, Maru D, Kopetz S, Navin NE. 2017. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res*. **27**(8):1287–1299.

Liu J, Dang H, Wang XW. 2018. The significance of intertumor and intratumor heterogeneity in liver cancer. *Exp Mol Med*. **50**(1): e416–e416.

Mai U, Sayyari E, Mirarab S. 2017. Minimum variance rooting of phylogenetic trees and implications for species tree reconstruction. *PLoS ONE*. **12**(8):e0182238.

Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, Baalsrud HT, Nederbragt AJ, Hanel R, Salzburger W, *et al.* 2016. Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet*. **48**(10):1204–1210.

Meijer A, van der Sanden S, Snijders BEP, Jaramillo-Gutierrez G, Bont L, van der Ent CK, Overduin P, Jenny SL, Jusic E, van der Avoort HGAM, *et al.* 2012. Emergence and epidemic occurrence of enterovirus 68 respiratory infections in the Netherlands in 2010. *Virology* **423**(1):49–57.

Miura S, Gomez K, Murillo O, Huuki LA, Vu T, Buturla T, Kumar S. 2018. Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics* **34**(23):4017–4026. doi:10.1093/bioinformatics/bty469

Nielsen R, Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain monte carlo approach. *Genetics* **158**(2):885–896.

Paradis E, Blomberg S, Bolker B, Brown J, Claude J, Sien Cuong H, Desper R, Didier G. 2019. Package 'ape'. *Anal Phylogenet Evol, Version*. **2**(4):47.

Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol Phylogenet Evol*. **28**(2): 171–185.

Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. 2015. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol*. **16**(1):91.

Pybus OG, Harvey PH. 2000. Testing macro–evolutionary models using incomplete molecular phylogenies. *Proc R Soc Lond Ser B: Biol Sci*. **267**(1459):2267–2272.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Syst Biol*. **67**(5):901.

Ross EM, Markowetz F. 2016. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome Biol*. **17**(1):1–14.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. *Genome Biol*. **14**(5):1–20.

Satas G, Zaccaria S, Mon G, Raphael BJ. 2020. Scarlet: single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Syst*. **10**(4):323–332.

Schwartz R, Schäffer AA. 2017. The evolution of tumour phylogenetics: principles and practice. *Nat Rev Genet*. **18**(4):213–229.

Stamatakis A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9): 1312–1313.

Sukumaran J, Holder MT. 2010. Dendropy: a python library for phylogenetic computing. *Bioinformatics* **26**(12): 1569–1571.

Tan G, Muffato M, Ledergerber C, Herrero J, Goldman N, Gil M, Dessimoz C. 2015. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. *Syst Biol*. **64**(5):778–791.

Tarabichi M, Salcedo A, Deshwar AG, Leathlobhair MN, Wintersinger J, Wedge DC, Van Loo P, Morris QD, Boutros PC. 2021. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat Methods* **18**(2):144–155.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect Math Life Sci*. **17**(2):57–86.

Tian Y, Kubatko L. 2017. Rooting phylogenetic trees under the coalescent model using site pattern probabilities. *BMC Evol Biol*. **17**(1):1–11.

Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. 2014. Efficient Bayesian inference under the structured coalescent. *Bioinformatics* **30**(16):2272–2279.

Wang L-G, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, *et al.* 2020. Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol Biol Evol*. **37**(2):599–603.

Wang Y, Navin NE. 2015. Advances and applications of single-cell sequencing technologies. *Mol Cell*. **58**(4):598–609.

Waskom ML. 2021. Seaborn: statistical data visualization. *J Open Source Softw*. **6**(60):3021.

Wickham H, 2016. *Elegant graphics for data analysis*. New York: Springer-Verlag.

Woodworth MB, Girskis KM, Walsh CA. 2017. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat Rev Genet*. **18**(4):230.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol*. **8**(1):28–36.

Zafar H, Navin N, Chen K, Nakhleh L. 2019. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Res*. **29**(11):1847–1859.

Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. 2017. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biol*. **18**(1):178.